HUMAN MUTATION 8:203–206 (1996)

SPECIAL ARTICLE

Mutation Nomenclature: Nicknames, Systematic Names, and Unique Identifiers

Ernest Beutler,* Victor A. McKusick, Arno G. Motulsky, Charles R. Scriver, and Franklin Hutchinson

Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California 92037 (E.B.), Center for Medical Genetics, Johns Hopkins Hospital, Baltimore, Maryland 21287-4922 (V.A.M.), Departments of Medicine and Genetics, Division of Medical Genetics, University of Washington, Seattle, Washington 98195 (A.G.M.), McGill University-Montreal Children's Hospital Research Institute, Montreal, Canada H3H 1P3 (C.R.S.), Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8040 (F.H.); Fax: 619-554-6927

Communicated by R.G.H. Cotton

INTRODUCTION

To enhance communication among scientists, they must use terminology that can be universally understood. Three levels of nomenclature are employed. Of these the most flexible is the group of terms that are designated *common* or *trivial* names. These are basically nicknames that are understood by those using them but that need to be defined in each document, because they do not unequivocally define what they represent. In contrast, *systematic* names do define exactly what they represent. When they are simple, systematic names are often used instead of trivial names. Third, with the development of interacting computer-based databases, it is desirable also to assign a unique identifier to each item. Such long numbers are difficult to remember, and although they may be cited in a document, they are not used to refer to the object.

Examples of this three-level hierarchy in nomenclature abound in science. Enzyme nomenclature is a familiar example. The enzyme known by the trivial names of glucose-6-phosphate dehydrogenase, G6PD, G6PDH, or Zwischenferment has the systematic name D-glucose-6-phosphate: NADP oxidoreductase and a unique identifier of E.C. 1.1.1.49. As another example, the trivial name LDH, lactic dehydrogenase, or lactate dehydrogenase denotes the enzymes L-lactate:NAD oxidoreductase with the unique identifier E.C. 1.1.1.27.

The detection of a large number of mutations has been possible for only a few years. Arriving at a consensus about nomenclature is not always achieved easily in new areas of research, because different terminologies spring up simultaneously in different laboratories as the field is developing.

COMMON (TRIVIAL) NAMES

Many kinds of trivial names have been used for the designation of human mutations. Largely because the first mutations, particularly those that affect hemoglobin, were designated by the changed amino acid, many investigators have used an amino acid-based notation for mutations. Beaudet and Tsui (1993) have put forward a system of trivial names based on amino acid nomenclature as a first step toward uniformity. Such a notation, which is actually phenotypic rather than genotypic, has the advantage of providing some information about the possible biologic effect of the mutation. In addition, the combination of two letters and a number is often easy to remember. Thus it has been useful as a trivial notation, assigning common names or nicknames to mutations. Others have used base numbers, either from cDNA or genomic DNA, as common names, since they more precisely define the actual mutation. Scientists need not be overly concerned about this diversity in designations, as long as they can agree on a systematic names and agree also to define the trivial name, if one is used, by the appropriate systematic name.

Received March 1, 1996; accepted June 12, 1996.

*To whom reprint requests/correspondence should be addressed.

This is manuscript 9857-MEM from The Scripps Research Institute.

SYSTEMATIC NAMES

The cardinal property of systematic names must be that by following easily understood rules, the systematic name unambiguously defines the object it represents. Other desirable properties include its compatibility with computerized systems and sufficient simplicity that it can be used as the trivial name if the scientist so desires.

Amino Acid-Based Nomenclatures

As noted, amino acid-based nomenclatures are widely used as trivial names. An excellent effort has been made to permit this nomenclature to also serve as a systematic nomenclature (AHCMN, 1996), but for the reasons cited below, an amino acid-based nomenclature is very difficult to adapt as systematic names:

1. Many amino acid changes can occur through several different base changes. Thus, whereas amino acid changes are predicted by base changes, base changes are not unambiguously predicted by amino acid changes.
2. There are at least three different starting points that are in common use in assigning codon numbers. In the older literature, when mutations were defined by protein sequencing, the initiator methionine was not counted. Thus, although the sickle mutation is commonly designated as E6V because of this historical fact, it would be designated E7V if discovered today. This convention has spilled over into the numbering of other mutant proteins, e.g., triosephosphate isomerase. Processed proteins add additional ambiguity. Sometimes numbering starts with the fully processed protein, sometimes with the partly processed protein, and sometimes with the native protein.
3. Insertions and deletions and mutations in promotors and introns cannot be incorporated into an amino acid-based system. Thus those who use an amino acid-based trivial nomenclature switch to nucleotide numbers when defining such events. This creates a system in which some numbers represent bases, some amino acids, and some introns. The convention states that names starting with a letter are amino acid names (e.g., R408W), using the one letter code for amino acids. Those beginning with a number are nucleotide names (e.g., 1348 C → T).

Although attempts have been made to adapt the amino acid-based system to such difficulties and thus to use it as a systematic nomenclature, the resulting set of rules become complex and the suggested designations cannot be sorted by a computer (Table 1).

TABLE 1. Examples of Mutations in the Cystic Fibrosis Transmembrane Protein Gene Using Nomenclature Proposed by Beaudet and Tsui (1993)[a]

D44G	1154insTC
A455E	2183del AA→G
S549R(A→C)	IVS4 + 1G→T
S549R(T→G)	IVS4-2A→C
Q39X	IVS19 + 10kbC→T
W1282X	R560Tsplice
Δ508	3120G→Asplice
1507del	M/V470
241delAT	1716G/A
852del22	125G/C

[a]Sorting these designations into a meaningful order by computer would be essentially impossible.

Nucleotide-based Nomenclatures

Nucleotide-based designations also have been used as trivial names for mutations and have the advantage of lending themselves more readily to a systematic nomenclature. Both cDNA-based and genomic-based numbering systems have been used.

cDNA-based Systems

cDNA-based systems lend themselves well to use as trivial names (Beutler, 1993) and are much more suitable for systematic names than is an amino-acid based nomenclature. However, there are certain problems. (1) cDNA-based designations cannot include introns using the same numbering system. As is the case with amino acid based systems, the introns must receive a separate designation, e.g., IVS2(+1)T, where the first nucleotide of the 5′ end of intron 2 is a T. (2) Some cDNAs are spliced differently in different tissues or even in the same tissue, and the start codon and the subsequent sequence are not always the same. This problem probably could be approached by using a reference cDNA, which might even be hypothetical and which contains all of the exons used in any tissue.

Genomic DNA-based Systems

A genomic-based system is without doubt the most robust basis of a systematic nomenclature. Such a system overcomes the difficulties imposed by introns and by alternative splicing. The A of the upstream ATG can be given the value of +1 as the beginning of the numbering system. The number −1 would be assigned to the base immediately 5′ to this A. Since polymorphisms can change the length of the genomic DNA, a refer-

ence sequence would need to be used, recognizing that in some instances the numbering for a given patient would not be correct for that individual. It would be clear, however, by comparison with the reference sequence exactly where the mutation was. It would not introduce any ambiguities. The only major difficulty inherent in such a system is that the entire genomic sequence often is not known when mutations begin to be described. This is probably only a temporary difficulty, as sequencing becomes easier and more and more sequences are deposited in the existing databases. This difficulty can be addressed by using cDNA numbers temporarily as the systematic nomenclature, prefacing the number with c. To prevent confusion, genomic numbers would be prefaced with g.

UNIQUE IDENTIFIERS

Fortunately, a system of unique identifiers for mutations has already been created in OMIM (McKusick, 1996). This system assigns numbers that have as an integer the locus number and assigns decimal numbers to mutations as they are discovered. It seems to us unwise to create a new system of identifiers, which would certainly not be unique, since the OMIM-based system already exists.

RECOMMENDATIONS FOR DISCUSSION

The following recommendations are based on the considerations listed above:

1. Trivial or common designations of mutations are at the discretion of the investigator. Amino acid-based designations, nucleotide-based designations, and other designations that have been used are all acceptable. Designations such as ΔF508 (cystic fibrosis), N370S, or 1226G (Gaucher disease), hemoglobin S, G6PD A^+, G6PD Mediterranean^{563T} are in the latter category. When amino acid-based nomenclature is used, the investigators are encouraged to use the system that has been proposed (AHCMN, 1996), since following specific rules confers on this nomenclature quasi-systematic properties. However, because the similar notation is used with different rules in the case of genes in which such a notation has been used extensively previously, it cannot be regarded as a truly systematic nomenclature. When such designations are used, therefore, they must be defined, when first used, by the systematic nomenclature. If the systematic name is used, which is encouraged, no further definition is required, but the inclusion of an OMIM unique identifier, if available, may be useful and is recommended.

2. Systematic names should be based upon nucleotide numbers. These numbers should be based on the genomic sequence whenever it is available. When the genomic sequence is not available, the cDNA sequences should be used. In both cases, numbering the positive series should start with the A of the upstream ATG; the negative series would describe the 5′ region of the gene. To avoid confusion, cDNA sequences should be preceeded by c and genomic sequences by a g. cDNA based sequences are always regarded as temporary, to be used only until the complete gene sequence is known. The reference sequence in the genomic databases should be identified. If only the cDNA sequence is available, mutations in introns should be designated IVSx, where x is the intron number and by numbering the nucleotides in introns +1, +2 . . . etc for the 5′ end of the intron and −1, −2 . . . etc for the 3′ end of the intron. No special notation is used when a mutation causes altered splicing, since splicing is often a stochastic process and it is difficult to include all the different possibilities that might be encountered in a notation that is intended primarily to unambiguously describe the mutation itself, not its consequences.

Certain rules are suggested to accommodate the many circumstances in which the mutation is not simply a single nucleotide change:

1. A nucleotide number followed by a base (A,C,G, or T) indicates that the nucleotide at that site is replaced by the designated nucleotide.

2. Deletions are designated by giving the nucleotide number(s) followed by *del*. When a deletion occurs in a repeating sequence, so that its actual location within the repeat is necessarily unknown, the most 3′ designation is arbitrarily assigned.

3. Insertions are designated by giving the nucleotide number before the insertion and the nucleotide number often with the base of bases inserted between the two numbers and the whole expression followed by *ins*. Thus insertion of AT after base 1273 would be written 1273 AT 1274 ins. When an insertion occurs in a repeating sequence, so that its actual location within the repeat is necessarily unknown, the most 3′ designation is arbitrarily assigned.

4. If there is more than one mutation on an allele, all should be included in the systematic nomenclature, separated by commas and enclosed in parentheses. However, polymorphic sites that are

TABLE 2. Examples of the Trivial (Common) Names, Systematic Names, and Unique Identifiers of Some Well Known Mutations

Gene	Trivial (common) names	Systematic names	Reference sequence	Unique identifier (McKusick, 1996)
CFTR	ΔF508	c1522–1524del	M28668	219700.0001
β globin	hemoglobin S	g20T	V00499	141900.0243
Glucose-6-P dehydrogenase	G6PD $A+^{376G}$ N126D	g10876G	X55448	305900.0001
Glucose-6-P dehydrogenase	G6PD $A-^{202A/376G}$ V68M,N126D	(g10153A,10876G)	X55448	305900.0002
Glucose-6-P dehydrogenase	G6PD Mediterranean^{563T} S188F	g12333T	X55448	305900.0006
Glucocerebrosidase	84GG	g451GG452 ins	J03059	230800.0014
Glucocerebrosidase	1226G or N370S	g5258G	J03059	230800.0003
Phenylalanine hydroxylase	R408W	c1222T	U49897	261600.0002

present in the normal population should not be included.

5. An inverted sequence is designated by *inv*. Thus if the nucleotides from 1234 to 1442 have been inverted, the designation would be 1234–1442inv.

6. No recommendations are made at this time for more complex re-arrangements. For such cases it is recommended that the sequence be deposited in an appropriate database and that the OMIM unique identifier be used.

Table 2 illustrates the trivial and proposed systematic names and the unique identifier for a number of mutations. An author discussing the common cystic fibrosis mutation might write "The ΔF508 (c1522–1524del, reference sequence M28668, Unique identifier 219700.0001) mutation is commonly encountered . . ." the first time the mutation is mentioned and then merely refer to it as ΔF508. For very common mutations such as this, referring to a report that gives the needed data would also suffice, e.g., "The ΔF508 mutation[87] is commonly encountered . . . ," where 87 is a reference in the bibliography that identifies the systematic name and the unique identifier of this mutation.

ACKNOWLEDGMENTS

The authors thank Drs. Arthur Beaudet and James Ostell for their valuable suggestions. This work was supported by the Stein Endowment Fund.

REFERENCES

Ad Hoc Committee on Mutation Nomenclature (AHCMN) (1996) Update on nomenclature for mutations. Hum Mutat 8:197–202 .

Beaudet AL, Tsui LC (1993) A suggested nomenclature for designating mutations. Hum Mutat 4:245–248.

Beutler E (1993) The designation of mutations. Am J Hum Genet 53:783–785.

McKusick VA (1996) http://www3.ncbi.nlm.nih.gov/omim/. World Wide Web.